

日 本 国 特 許 庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日 2 0 0 1 年 6 月 1 8 日
Date of Application:

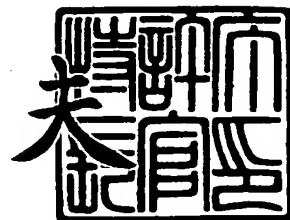
出 願 番 号 特 願 2 0 0 1 - 1 8 3 8 5 6
Application Number:
[ST. 10/C]: [J P 2 0 0 1 - 1 8 3 8 5 6]

出 願 人 科学技術振興事業団
Applicant(s):

2 0 0 3 年 1 0 月 2 2 日

特許庁長官
Commissioner,
Japan Patent Office

今 井 康



【書類名】 特許願

【整理番号】 A000103431

【特記事項】 特許法第 3 0 条第 1 項の規定の適用を受けようとする特
許出願

【提出日】 平成13年 6月18日

【あて先】 特許庁長官 殿

【国際特許分類】 G06F 17/00

【発明の名称】 相同性解析システム、相同性解析方法及び相同性解析プ
ログラム

【請求項の数】 11

【発明者】

【住所又は居所】 埼玉県熊谷市久下 2 0 3 8 - 3

【氏名】 篠沢 隆雄

【発明者】

【住所又は居所】 群馬県桐生市西久方町 1 - 5 - 2 3 パークハイツ 3 0
6

【氏名】 堀池 徳祐

【特許出願人】

【識別番号】 396020800

【氏名又は名称】 科学技術振興事業団

【代理人】

【識別番号】 100058479

【弁理士】

【氏名又は名称】 鈴江 武彦

【電話番号】 03-3502-3181

【選任した代理人】

【識別番号】 100092196

【弁理士】

【氏名又は名称】 橋本 良郎

【選任した代理人】

【識別番号】 100091351

【弁理士】

【氏名又は名称】 河野 哲

【選任した代理人】

【識別番号】 100088683

【弁理士】

【氏名又は名称】 中村 誠

【手数料の表示】

【予納台帳番号】 011567

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 相同性解析システム、相同性解析方法及び相同性解析プログラム

【特許請求の範囲】

【請求項 1】 解析対象となる解析対象データ群が、前記解析対象データ群とは異なる第 1 のデータ群あるいは第 2 のデータ群のいずれに類似するかを解析するための相同性解析システムであって、

前記解析対象データ群と第 1 のデータ群のそれぞれのデータ群に含まれるデータの相同性を示す第 1 の相同値 x を算出するものであって、前記しきい値 E を n 個設定して各しきい値 E_i ($i = 1, 2, \dots, n$) 毎に第 1 の相同値 x_i を算出する第 1 の相同値算出手段と、

前記解析対象データ群と第 2 のデータ群のそれぞれのデータ群に含まれるデータの相同性を示す第 2 の相同値 y を算出するものであって、前記しきい値 E を n 個設定して各しきい値 E_i ($i = 1, 2, \dots, n$) 毎に第 2 の相同値 y_i を算出する第 2 の相同値算出手段と、

第 1 の相同値 x_i 及び第 2 の相同値 y_i 並びにしきい値の数 n との関係に基づいて、前記解析対象データ群が第 1 のデータ群又は第 2 のデータ群のいずれに類似するかを判定する相同性決定手段と

を具備してなることを特徴とする相同性解析システム。

【請求項 2】 前記第 1 の相同値算出手段は、

前記解析対象データ群と第 1 のデータ群のそれぞれのデータ群に含まれるデータの相同性が所定のしきい値 E 以上である場合には相同性有りと判定し、相同性有りとなるデータの数を第 1 の相同値 x として算出し、

前記第 2 の相同値算出手段は、

前記解析対象データ群と第 2 のデータ群のそれぞれのデータ群に含まれるデータの相同性が所定のしきい値 E 以上である場合には相同性有りと判定し、相同性有りとなるデータの数を第 1 の相同値 y として算出する

ことを特徴とする請求項 1 に記載の相同性解析システム。

【請求項 3】 前記第 1 のデータ群は n_A 個のデータを有し、前記第 2 のデ

ータ群は n_B 個のデータを有する場合、

前記第 1 の相同値算出手段は、一つのしきい値 E_i に対して前記第 1 のデータ群の各データ毎に第 1 の相同値 x_{ij} ($j = 1, 2, \dots, n_A$) を算出し、その算出されたしきい値 E_i についての第 1 の相同値 x_i の平均値 \bar{x}_i を算出し、

前記第 2 の相同値算出手段は、一つのしきい値 E_i に対して前記第 2 のデータ群の各データ毎に第 1 の相同値 y_{ij} ($j = 1, 2, \dots, n_B$) を算出し、その算出されたしきい値 E_i についての第 2 の相同値 y_i の平均値 \bar{y}_i を算出し、

前記相同性決定手段は、

【数 1】

$$u_i = \sqrt{\frac{1}{n_A + n_B - 2} \left\{ \sum_{j=1}^{n_A} (x_{ij} - \bar{x}_i)^2 + \sum_{k=1}^{n_B} (y_{ik} - \bar{y}_i)^2 \right\}}$$

とした場合に、以下の等式

【数 2】

$$Z_i^{(1)} = \frac{\bar{x}_i - \bar{y}_i}{u_i} \cdot \sqrt{\frac{n_A \cdot n_B}{n_A + n_B}} \quad (i = 1, 2, \dots, n)$$

により、第 1 のデータ群あるいは第 2 のデータ群のいずれに類似するかを示す相同性決定値 $Z_i^{(1)}$ を算出することを特徴とする請求項 1 に記載の相同性解析システム。

【請求項 4】 前記相同性決定値 $Z_i^{(1)}$ が t 分布に従い、自由度 α とすると、 $t_{\alpha}(0, 10)$ よりも大きい場合には前記解析対象データ群は前記第 1 のデータ群と相同性を持つデータ数が多いと判定する判定結果導出手段をさらに備えることを特徴とする請求項 3 に記載の相同性解析システム。

【請求項 5】 前記相同性決定値 $Z_i^{(1)}$ が t 分布に従い、自由度 α とすると、 $-t_{\alpha}(0, 10)$ よりも小さい場合には前記解析対象データ群は前記第 2 のデータ群と相同性を持つデータ数が多いと判定する判定結果導出手段をさらに備えることを特徴とする請求項 2 に記載の相同性解析システム。

【請求項 6】 前記判定結果導出手段はさらに、 s を $Z_i^{(1)}$ の標準偏差、 $\bar{Z}_i^{(1)}$ を $Z_i^{(1)}$ の平均値とすると、

【数 3】

$$Z^{(2)} = \frac{|\overline{Z^{(1)}}| - t_{n_A+n_B-2}(0.10)}{s/\sqrt{n-1}}$$

の等式で示される相同性可否決定値 $Z^{(2)}$ を算出し、該相同性可否決定値 $Z^{(2)}$ が所定の値 $t_{n-1}(0, 10)$ 未満の場合には相同性決定値 $Z_i^{(1)}$ が有効な値でないと判定することを特徴とする相同性可否判定手段を備えることを特徴とする請求項 4 又は 5 に記載の相同性解析システム。

【請求項 7】 前記自由度 α は、 $n_A + n_B - 2$ であることを特徴とする請求項 4 乃至 6 に記載の相同性解析システム。

【請求項 8】 前記第 1 及び第 2 の相同値算出手段は、BLAST 法により相同値 x_i 及び y_i を算出することを特徴とする請求項 1 乃至 7 に記載の相同性解析システム。

【請求項 9】 前記解析対象データ群、第 1 のデータ群及び第 2 のデータ群は、遺伝子の配列を示すデータであることを特徴とする請求項 1 乃至 8 に記載の相同性解析システム。

【請求項 10】 解析対象となる解析対象データ群が、前記解析対象データ群とは異なる第 1 のデータ群あるいは第 2 のデータ群のいずれに類似するかを解析するための相同性解析方法であって、

前記解析対象データ群と第 1 のデータ群のそれぞれのデータ群に含まれるデータの相同性を示す第 1 の相同値 x を算出するステップであって、この第 1 の相同値 x を、前記しきい値 E を n 個設定して各しきい値 E_i ($i = 1, 2, \dots, n$) 毎に第 1 の相同値 x_i として算出するステップと、

前記解析対象データ群と第 2 のデータ群のそれぞれのデータ群に含まれるデータの相同性を示す第 2 の相同値 y を算出するステップであって、この第 2 の相同値 y を、前記しきい値 E を n 個設定して各しきい値 E_i ($i = 1, 2, \dots, n$) 毎に第 2 の相同値 y_i として算出するステップと、

第 1 の相同値 x_i 及び第 2 の相同値 y_i 並びにしきい値の数 n との関係に基づいて、前記解析対象データ群が第 1 のデータ群又は第 2 のデータ群のいずれに類似

するかを判定するステップと

を有することを特徴とする相同性解析システム。

【請求項 11】 解析対象となる解析対象データ群が、前記解析対象データ群とは異なる第 1 のデータ群あるいは第 2 のデータ群のいずれに類似するかを解析するための相同性解析プログラムであって、コンピュータを、

前記解析対象データ群と第 1 のデータ群のそれぞれのデータ群に含まれるデータの相同性を示す第 1 の相同値 x を算出するものであって、前記しきい値 E を n 個設定して各しきい値 E_i ($i = 1, 2, \dots, n$) 毎に第 1 の相同値 x_i を算出する第 1 の相同値算出手段と、

前記解析対象データ群と第 2 のデータ群のそれぞれのデータ群に含まれるデータの相同性を示す第 2 の相同値 y を算出するものであって、前記しきい値 E を n 個設定して各しきい値 E_i ($i = 1, 2, \dots, n$) 毎に第 2 の相同値 y_i を算出する第 2 の相同値算出手段と、

第 1 の相同値 x_i 及び第 2 の相同値 y_i 並びにしきい値の数 n との関係に基づいて、前記解析対象データ群が第 1 のデータ群又は第 2 のデータ群のいずれに類似するかを判定する相同性決定手段

として機能させることを特徴とする相同性解析プログラム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、複数のデータからなるデータ群同士の相同性を解析するための相同性解析システム、相同性解析方法及び相同性解析プログラムに関する。

【0002】

【従来の技術】

遺伝子の類似性を比較する従来の方法として最も普及している方法は、「系統樹解析」である。この方法は、遺伝子の類似性を含めた相互関係を最も厳密に解析することができる。しかし、多数の遺伝子に関する解析を行う場合は、それに要する時間が長い点において大きな欠点を要する。また、生物の固体間の全体的な類似性、すなわち多数の遺伝子のマスとしての相関関係などの解析には、採用

した遺伝子が本当にその生物間の関係を代表しているものであるかという点で、適当な遺伝子を選んでいとは限らない。遺伝子の相同性・類似性を比較する方法としては、上記の系統樹解析に加えて、BLAST法やFASTA法が代表的である。ここで、BLAST法とは、遺伝子のデータベースの中からある特定の遺伝子と類似性を有する遺伝子を選別するプログラムを用いて解析を行う方法である。

【0003】

【発明が解決しようとする課題】

上記したようなBLAST法やFASTA法では、遺伝子間の類似性の度合い、すなわちどれほど類似性が高いかに応じて、どのような遺伝子がどの程度の類似性があるかということを解析し、遺伝子を選別・特定することができる。しかし、これらの方法は類似性の度合いを設定する領域、すなわちしきい値を固定しているために、他の類似性の度合いにおいて、どのような遺伝子が存在するかなどの解析を行う点で不十分であった。

【0004】

本発明は上記課題を解決するためになされたもので、その目的とするところは、多数の遺伝子同士を定量的にかつ正確に比較することが可能な相同性解析システム、相同性解析方法及び相同性解析プログラムを提供することにある。

【0005】

【課題を解決するための手段】

この発明のある観点によれば、解析対象となる解析対象データ群が、前記解析対象データ群とは異なる第1のデータ群あるいは第2のデータ群のいずれに類似するかを解析するための相同性解析システムであって、前記解析対象データ群と第1のデータ群のそれぞれのデータ群に含まれるデータの相同性を示す第1の相同値 x を算出するものであって、前記しきい値 E を n 個設定して各しきい値 E_i ($i=1, 2, \dots, n$) 毎に第1の相同値 x_i を算出する第1の相同値算出手段と、前記解析対象データ群と第2のデータ群のそれぞれのデータ群に含まれるデータの相同性を示す第2の相同値 y を算出するものであって、前記しきい値 E を n 個設定して各しきい値 E_i ($i=1, 2, \dots, n$) 毎に第2の相同値 y_i を算出する第

2の相同値算出手段と、第1の相同値 x_i 及び第2の相同値 y_i 並びにしきい値の数 n との関係に基づいて、前記解析対象データ群が第1のデータ群又は第2のデータ群のいずれに類似するかを判定する相同性決定手段とを具備してなることを特徴とする相同性解析システムが提供される。

【0006】

このような構成によれば、多数のデータからなるデータ群同士の相同性、すなわち類似性をまとめて評価できる。また、データとして遺伝子データを適用することにより、多数の遺伝子群の相同性をまとめて評価できる。よって、従来のように系統樹を描くために代表する遺伝子を選択する必要が無くなるため、遺伝子群としての特徴をより正確に理解することが可能となる。

【0007】

例えば、従来は多数の遺伝子群間の関係を調べる方法では、その遺伝子群を代表する遺伝子を数個選び出し、その系統樹を描く方法であった。しかし、この方法では選び出した遺伝子が本当に代表として適当であるか判別することができず、選び出した遺伝子が適切か否かで結果が大きく異なっていた。これに対して、対象とするデータ群であるすべての遺伝子（ORF）について相同性検索を行い、あるしきい値を達している遺伝子（ORF）数を数えるBLAST法やFASTA法によれば、互いの関係を推測することができる。しかしながら、この方法でも、遺伝子群にいくつかの異なった由来の遺伝子群が混じっていた場合には、しきい値の基準を変えると結果がかなり異なってしまう特徴を有していた。そこで、本発明のように、いくつものしきい値で相同性検索を行い、それぞれのしきい値において相同性のある遺伝子（ORF）の数をプロットし、遺伝子群間の関係を推測する方法を取ることで、しきい値の適切性を考慮することなく極めて安定した相同性評価結果を得ることができる。

【0008】

現在、ヒトゲノム解明以後の大きな流れとして、病気の原因となる遺伝子の選別特定が重要となってきた。候補となった遺伝子を実際のヒトDNA試料からPCR遺伝子増幅法により、DNA断片として作成し、大腸菌などの発現系でその遺伝子が合成するタンパク質を実際に作らせ、その活性をみることや、ある

いはそのタンパク質に対する抗体を作成することが行われる。抗体は実際にヒトの病理切片などで試すことで病気の検出方法などに有用である。また、病気の原因となる遺伝子の特定は患者の組織の病域の特定、あるいは健常人の遺伝子を解析することで、その病気にかかりやすい人を選別することができる。そうすることで、患者に対する治療の方法、例えば、保因者とそうでない人に対しては、例えばポリプ切除の緊急性の判断に応用できる。健常人に関しては、その病気に関する診断検査のそれぞれの人の遺伝子情報に基づき使い分けることが可能である。

【0009】

また、システム（装置）に係る本発明は、その装置により実現される方法の発明としても成立する。

【0010】

また、装置または方法に係る本発明は、コンピュータに当該発明に相当する手順を実行させるための（あるいはコンピュータを当該発明に相当する手段として機能させるための、あるいはコンピュータに当該発明に相当する機能を実現させるための）プログラム、このプログラムを記録したコンピュータ読取り可能な記録媒体としても成立する。

【0011】

【発明の実施の形態】

本発明の骨子は、解析対象となるデータ群（解析対象データ群）と、他のデータ群との相同性を算出し、解析対象データ群がいずれのデータ群と相同性を有するか否かを判定する点にある。

【0012】

以下、図面を参照しながら本発明の一実施形態を説明する。

【0013】

図1は本発明の第1実施形態に係る遺伝子解析システム10の全体構成を示す図である。図1に示すように、プロセッサ1と、このプロセッサ1に接続された出力手段2と、プロセッサ1に接続されたアミノ酸配列記憶手段3と、相同値記憶手段4から構成される。プロセッサ1は、BLAST解析手段11とT検定手

段 12 を備える。

【0014】

プロセッサ 1 が図示しない記録媒体から所定のプログラムを読み出すことにより各手段 11 及び 12 としてプロセッサ 1 が機能する。BLAST 解析手段 11 は、複数のデータからなるあるデータ群と他のデータ群の相同性を相同値として算出する相同値算出手段を備える。また、T 検定手段 12 は、相同性決定値算出手段、相同性可否決定値算出手段及び判定結果導出手段を備える。相同性決定値算出手段は、相同性、すなわちあるデータ群と他のデータ群の類似性を数値により示した相同性決定値 $Z_i(1)$ を算出する。相同性可否決定値算出手段は、相同性決定値算出手段で算出された相同性決定値 $Z_i(1)$ が相同性を示す値として適切であるか否かを判定するための数値である相同性可否決定値 $Z(2)$ を算出する。判定結果導出手段は、相同性決定値算出手段で算出された相同性決定値 $Z_i(1)$ と相同性可否決定値算出手段で算出された相同性可否決定値 $Z(2)$ に基づいていずれのデータ群に相同性の多いデータを有するかの判定を行う。

【0015】

次に、図 2 を用いて本実施形態に係る遺伝子解析に必要な遺伝子データを取得するプロセスを説明する。

【0016】

まず、インターネットなどのネットワークを介して、遺伝子解析システム 10 の BLAST 解析手段 11 が酵母 ORF (Open Reading Frames) アミノ酸配列を取得する (s21)。次に、この取得したアミノ酸配列を機能別 43 のカテゴリーに分類する (s22)。そして、得られた機能別のカテゴリーから、ミトコンドリアからの遺伝子移行の影響を受けていない酵母 ORF を抽出する (s23)。なお、原核生物として、細胞寄生性の原核生物を除く原核生物の ORF を 15 種類取得する。

【0017】

以上により、真核生物である酵母の ORF と、原核生物の 15 種類の ORF が得られる。得られた各 ORF は、アミノ酸配列記憶手段 3 に格納される。なお、これら各 ORF (遺伝子データ) は、例えば酵母であれば酵母全体の各 ORF が

ひとまとまりとして遺伝子データ群として格納され、例えば原核生物であれば、ある原核生物全体の各ORF（遺伝子データ）がひとまとまりとして遺伝子データ群として格納される。また、各ORFはアミノ酸配列により特定される。このアミノ酸配列記憶手段3に格納された遺伝子データ群の概念図を図3に示す。図3では、酵母ORFは全部で p 個、古細菌AのORFは n_1 個、古細菌BのORFは n_2 個、真正細菌CのORFは n_3 個、真正細菌DのORFは n_4 個示してある。

【0018】

次に、BLAST解析手段11は、以上のようにして得られた16種類のORFのうち、解析対象として酵母のORFを選択し、この解析対象である酵母ORFと15種類のORFのそれぞれとの相同値ヒット数（オルソログス遺伝子数）をBLAST法を用いて算出する。

【0019】

オルソログス遺伝子数を検出するには、まず酵母の全ORFと各細菌の全ORFに対してORF間の相同性の指標であるE-値（相同性と負の相関）を算出する。酵母と各細菌でお互いに最小のE-値を持つORFをオルソログス遺伝子と判定する。オルソログス遺伝子の数（ヒット数）が生物間の相同性を表す。そして、様々なしきい値（E-値）に対してこのヒット数を算出する。

【0020】

図4はBLAST法の概念図である。図4に示すように、酵母のORFを細胞寄生性の原核生物を除く15種類の原核生物のORFに対して相同性検索を行う。相同性検索は、まずプロセッサ1がアミノ酸配列記憶手段3から酵母のORF1と古細菌A₁のORFを読み出す。そして、相同性の指標であるE-値を算出し、最小のE-値を持つORFを古細菌A₁の酵母ORF1との第一ヒットORFとし、そのORF名とその時のE-値を記憶させる。これを全ORFに対して繰り返す。他の古細菌A₂, ..., A_n、真正細菌に対しても同様に行う。BLAST法については、例えば2001年2月1日発行の“Nature Cell Biology, Vol.3, No.2, pp210-214”に開示されている。また、遺伝子を比較する方法はBLAST法以外でも、アミノ酸配列同士、塩基配列同士、分子レベルなど、遺伝子を情報

として特定できるいかなるレベルで比較してもよい。

【0021】

さらに、図5はオルソログス遺伝子数で表される生物間の相同性の判定の概念を示す図である。図5では、矢印の先は第一ヒット遺伝子である。但し、互いのE値はしきい値を超えている。丸印で示したのは、互いにオルソログス遺伝子であることを示している。一方、酵母ORF2にとっての第一ヒット遺伝子は古細菌A₁のORF3であり、古細菌A₁のORF3にとっての第一ヒット遺伝子は酵母ORF3である。従って、酵母ORF2の第一ヒット遺伝子と古細菌A₁のORF3の第一ヒット遺伝子は一致しない。オルソログス遺伝子の数が酵母と古細菌A₁との生物間の相同性を表す。

【0022】

図5のように、酵母と古細菌A₁に着目すると、この酵母のORF1の古細菌A₁との第一ヒットORF-aとそのときのE-値を読み出し、そして古細菌A₁のORF-aの酵母との第一ヒットORF- α とそのときのE-値を読み出す。ORF1とORF- α が一致し、両E-値がしきい値Eより大きいとき、オルソログス遺伝子と判定し、生物間相同値ヒット数としてカウントする。次に、酵母のORF2の古細菌A₁との第一ヒットORF-bとそのときのE-値を読み出し、同様に古細菌A₁とのオルソログス遺伝子の有無を判定する。ORF-bがオルソログス遺伝子のとき、生物間相同値ヒット数としてカウントする。

【0023】

このようにして、酵母ORF3～酵母ORF_pについても同様にオルソログスORFの有無を判定する。そして、得られたカウント値を古細菌A₁との相同値ヒット数xとして相同値記憶手段4に格納する。

【0024】

以上は、酵母と古細菌A₁に着目して説明したが、同様に酵母と古細菌A₂, ..., A_nについても行い、それぞれを酵母との相同値ヒット数xとしてそれぞれ相同値記憶手段4に格納する。

【0025】

また、プロセッサ1のBLAST解析手段11は、古細菌A₁, ..., A_nについ

ての相同値ヒット数の平均値 \bar{x} を算出し、得られた平均値を相同値記憶手段 4 に格納する。

【0026】

また、このような相同値ヒット数の算出を他の古細菌遺伝子データ群や、真正細菌遺伝子データ群などについても行う。

【0027】

また、 p 個の遺伝子数である酵母遺伝子データ群と n_A 個の古細菌の遺伝子データ群とのオルソログス遺伝子検出において、あるしきい値 E_i について、 j 番目の古細菌について得られた相同値ヒット数を x_{ij} とする。

【0028】

また、 p 個の遺伝子数である酵母遺伝子データ群と n_B 個の真正細菌の遺伝子データ群とのオルソログス遺伝子検出において、あるしきい値 E_i について、 k 番目の真正細菌について得られた相同値ヒット数を y_{ik} とする。

【0029】

以上のようにして得られた相同値ヒット数としきい値 E との関係を示したのが図 6 である。横軸は $-\log E$ 、縦軸は相同値ヒット数である。なお、この図 6 では、それぞれの折れ線グラフが酵母遺伝子と比較する各細菌に対応している。類似性の度合いを厳しく、すなわちしきい値 E を小さくしていくにつれ、条件を満たす遺伝子数、すなわち相同値ヒット数は漸減しているのが分かる。

【0030】

以上により取得された各種相同値ヒット数に基づき、プロセッサ 1 の T 検定手段は、以下による T 検定処理を行う。

【0031】

T 検定処理は、相同性決定値 $Z_i^{(1)}$ の算出、相同性可否決定値 $Z^{(2)}$ の算出及び判定結果導出の 3 つの処理からなる。

【0032】

まず、相同性決定値 $Z_i^{(1)}$ の算出手法について酵母遺伝子データ群が古細菌遺伝子データ群と真正細菌遺伝子データ群のいずれに相同性が高い遺伝子を多く共有するかを解析する場合について説明する。

【0033】

一例として、上述した各しきい値 E_i ($i = 1, 2, \dots, n$) において、有意水準 5 % の片側 T 検定をヒット数が少なくとも 5 以上という基準を満たす領域で行う。

【0034】

そして、相同性決定値 $Z_i^{(1)}$ (1) を以下の等式により算出する。

【0035】

【数4】

$$Z_i^{(1)} = \frac{\bar{x}_i - \bar{y}_i}{u_i} \cdot \sqrt{\frac{n_A \cdot n_B}{n_A + n_B}} \quad (i = 1, 2, \dots, n)$$

【0036】

なお、以上に示す等式で、 \bar{x}_i は i 番目の E 値でのヒット数 x_{ij} ($j = 1, 2, \dots, n_A$) の平均値を、 \bar{y}_i は i 番目の E 値でのヒット数 y_{ik} ($k = 1, 2, \dots, n_B$) の平均値を示す。 n は、BLAST法で解析した際のしきい値の数を示している。また、不偏分散 u_i は以下の等式で示される。

【0037】

【数5】

$$u_i = \sqrt{\frac{1}{n_A + n_B - 2} \left\{ \sum_{j=1}^{n_A} (x_{ij} - \bar{x}_i)^2 + \sum_{k=1}^{n_B} (y_{ik} - \bar{y}_i)^2 \right\}}$$

【0038】

以上のようにして得られた相同性決定値 $Z_i^{(1)}$ (1) を統計処理したのが図7である。図7では、横軸に $-\log E$ を、縦軸に相同性決定値 $Z_i^{(1)}$ (1) をとってあり、図6に示されるデータに基づいて得られた値である。相同性有りとは判断された遺伝子数を 5 以上の領域で棒グラフにより示している。相同性決定値 $Z_i^{(1)}$ (1) が $t_{n_A+n_B-2}(0, 10)$ ($= 1.771$) 以上又は $-t_{n_A+n_B-2}(0, 10)$ ($= -1.771$) 以下の場合には有意水準 5 % (2つの母平均が等しいと仮定したとき、標本平均が等しくない確率が 5 % 以下) で、各々のしきい値 E に対し

て古細菌あるいは真正細菌の遺伝子群が出芽酵母の遺伝子群と類似性が高い。

【0039】

次に、相同性可否決定値 $Z(2)$ の算出手法について説明する。この相同性可否決定値 $Z(2)$ の算出は、1 回目の T 検定、すなわち相同性決定値 $Z_i(1)$ の算出における相同性決定値 $Z_i(1)$ が、 $t_{nA+nB-2}(0, 10)$ ($=1.771$) より設定した E_i 全体に対して優位に大きいか、 $-t_{nA+nB-2}(0, 10)$ ($=-1.771$) より小さいか、あるいはどちらの傾向も無いかを判定するために行われる。そのため、自由度 $n-1$ において以下の等式により、相同性可否決定値 $Z(2)$ と $t_{n-1}(0, 10)$ を計算する。

【0040】

【数6】

$$Z^{(2)} = \frac{|\bar{Z}^{(1)}| - t_{nA+nB-2}(0.10)}{s/\sqrt{n-1}}$$

【0041】

ここで、 s は相同性決定値 $Z_i(1)$ の標準偏差、 $Z_i(1)$ は $Z_i(1)$ の平均値である。相同性可否決定値 $Z(2)$ が基準値 t_{n-1} より大きいときは $Z_i(1)$ が $t_{nA+nB-2}$ よりも有意水準 5% で大きいと判定できる。

【0042】

次に、以上により算出された相同性決定値 $Z_i(1)$ 及び相同性可否決定値 $Z(2)$ に基づいて、判定結果導出を行う。判定結果導出は、図 8 に示す判定テーブルを用いて行う。図 8 に示すように、 $Z(2) / t_{n-1}(0, 10)$ が 1 以上で、 $Z_i(1)$ の平均値が正であれば、そのカテゴリーにおける酵母 ORF 群は真正細菌よりも古細菌との方に多くの相同性を有する ORF を含むと判定できる。また、 $Z(2) / t_{n-1}(0, 10)$ が 1 以上で、 $Z_i(1)$ の平均値が負であれば、そのカテゴリーにおける酵母の ORF 群は古細菌よりも真正細菌との方に多くの相同性を有する ORF を含むと判定できる。また、 $Z(2) / t_{n-1}(0, 10)$ が 1 より小さい場合はいずれの細菌に相同性を有するかを判定できないと判定する。

【0043】

このような構成によれば、多数の遺伝子群の相同性をまとめて評価できる。よって、従来のように系統樹を描くために代表する遺伝子を選択する必要がなくなるため、遺伝子群としての特徴をより正確に理解することが可能となる。

【0044】

例えば、この方法を用いて真核生物の起源を探ることができる。従来は、主に rRNA と小数のタンパク質について系統樹を作成し、推測していた。そして、その推測によれば、DNA複製、転写、翻訳等の遺伝子群は古細菌起源で、エネルギー代謝の遺伝子群はミトコンドリアの共生が起源で、その他の遺伝子群は古細菌と真正細菌のモザイク構造になっていると予想されていた。また、その他説が分かれていた。

【0045】

そこで、酵母の持つそれぞれの機能の由来を考察するため、細胞寄生性細菌を除く15種のバクテリアのORF (Open Reading Frames) を、機能別に分類された酵母のORFと比較する。すなわち、上述したBLAST解析、T検定の解析対象を酵母遺伝子データ群、比較対照を古細菌遺伝子データ群及び真正細菌遺伝子データ群とする。その結果、43の機能別に分類された酵母の遺伝子群のうち、20の遺伝子群は古細菌、又は真正細菌のどちらか一方に相同性の高い遺伝子を多く含むことが分かった。一般に、進化系統の近い生物間では相同性の高い遺伝子を多く保持していることが知られていることから、酵母遺伝子の機能カテゴリー別にそれぞれの起源を推定できる。

【0046】

具体的にこのBLAST解析及びT検定を行うことにより、DNA合成及び複製、転写、翻訳、減数分裂、細胞周期調節、小胞体形成、核形成などの遺伝子群は古細菌ORFと、エネルギー代謝、各種物質代謝、細胞内への物質輸送、ストレス応答、解毒、イオンホメオスタシスの遺伝子群は真正細菌ORFと、それぞれ相同性が高い遺伝子を多く共有することが分かった。

【0047】

従って、核（遺伝情報）に関連した遺伝子は古細菌の遺伝子が、また細胞質（ホメオスタシス）に関連した遺伝子は真正細菌の遺伝子が起源であると考えられ

る。また、このことは、どちらかの遺伝子群が後から他方の生物に置き換わったことを示す。

【0048】

これらの結果は、真核生物の核は古細菌の真正細菌への共生に由来することを示唆する。すなわち、相同性判定により遺伝子群間の関係を推測することにより、真核生物は真正細菌へ古細菌が共生することによって誕生したと考えられる可能性が高い。

【0049】

このことを現在有力とされる各説に応じて考えると、以下のようになる。

【0050】

まず、真正細菌由来の遺伝子はミトコンドリア共生の結果もたらされたのではないかとする説に対しては、ミトコンドリア関連遺伝子は解析のデータから削除されており、影響しないと考えられると結論づけられる。

【0051】

また、遺伝子水平伝達の蓄積ではないかとする説に対しては、遺伝子水平伝達は、偶然の要素によって起こる。特定の機能のみ遺伝子水平伝達が起こり、その機能の遺伝子群がそっくり入れ替わるとは考えにくいと結論づけられる。

【0052】

また、古細菌が真正細菌に進入し、共生した結果ではないかとする説に対しては、ある特定の機能に関する遺伝子群が置き換わったということは、その遺伝子群だけが置き換わりやすい特殊な環境であったことを示す。細胞内共生による核形成のプロセスはこの条件を満たし、またミトコンドリアや葉緑体の進化的形成が細胞内共生によっていることが広く認知されていることからもっともらしいといえると結論付けられる。従って、古細菌が真正細菌に進入し、共生した結果、真核生物が生じたと考えられる。

【0053】

具体的には、真正細菌に細胞内共生した古細菌は真正細菌が合成する代謝産物を利用し続けるにつれ、それらの合成に必要な遺伝子群を失った。一方真正細菌は細胞周期などを古細菌に支配され、最終的には遺伝子複製からタンパク質合成

までのプロセスも古細菌に依存し、それ以外の遺伝子は徐々に移行したと推測できる。このことを模式的に示したのが図9である。図9に示すように、細胞周期調節、核形成、DNA複製、RNA転写、小胞体形成、翻訳、リボソーム形成などに関連する核関連の遺伝子91は、古細菌遺伝子が起源であると考えられ、細胞への物質輸送、代謝、ストレス応答、解毒、イオンホメオスタシスなどに関連する細胞質92は、真正細菌遺伝子を起源とする遺伝子群に関連する。また、ミトコンドリア形成、ミトコンドリア輸送などに関連するミトコンドリア93は、ミトコンドリアの共生に由来していると考えられる。

【0054】

このように、遺伝子群をドメインレベルにまで細分化することによる相同性解析の検出能の上昇と、祖先型遺伝子をタンパク質として再現し、その活性を確認することで、遺伝子の由来に関する推測の信頼性が格段に向上する。

【0055】

以上説明したように、本実施形態によれば、全ゲノム配列が明らかになっている酵母とバクテリアのORFデータを用い、真核生物、真正細菌、古細菌の3つの生物群における進化的な関係を明らかにすることができる。また、ヒトやその他の生物のORFデータを用いてその相同性を解析することにより、あらゆる生物同士の相同性をしきい値の設定条件に依存せずに安定して極めて簡便に判定することができる。

【0056】

本発明は上記実施形態に限定されるものではない。BLAST法を用いて相同値ヒット数を算出したが、他の相同値算出法を適用することも可能であることはもちろんである。

【0057】

また、例えば、非常に多数の遺伝子集団を用いて、ある目的に従って比較したり、特定の遺伝子を選別したり、あるいは選別された遺伝子を具体的に特定することに本発明を適用することができる。すなわち、対象とする遺伝子グループが複数ある場合には、それぞれのしきい値（E値）で計算することにより、非常に多数の遺伝子を有する複数のグループ間におけるそれぞれの遺伝子の、ある特徴

を有する遺伝子群との相同性類似性を比較すること、さらにそこで対象となった遺伝子を抽出、特定することを非常に簡便に行うことができる。さらに、対象とした遺伝子群の中に、そのグループの中では特異な特徴を有する遺伝子を特定することもできる。例えば、生物の進化過程においては親から子への伝達ではない、いわゆる水平伝達で導入された遺伝子の抽出にも適用可能である。

【0058】

特定された遺伝子を解析した生物の中から取り出し、DNA塩基配列を知ること、あるいは類似の遺伝子を検索し、類似の遺伝子の情報に基づきそれらの遺伝子の機能を推測することが可能となる。また、ある目的に従った人材の選別などにも応用できる。すなわち、非常に多数のデータを含むグループ（データ群）の中から、ある一定の条件を満たす、又はその条件の度合いを変化させつつデータを選別、特定できる。グループ間で相互に相同性を解析することにより、お互いの条件を満たすペアを選別することもできる。

【0059】

また、現在、ヒトゲノム解明以後の大きな流れとして、病気の原因となる遺伝子の選別特定が重要となっている。候補となっている遺伝子を実際のヒトDNA試料からPCR遺伝子増幅法により、DNA断片として作成し、大腸菌等の発現系でその遺伝子が合成するタンパク質を実際に作らせ、その活性をみることや或いはそのタンパク質に対する抗体を作成することが行われる。抗体は実際にヒトの病理切片などで試すことで病気の検出方法等に有用である。また、病気の原因となる遺伝子の特定は患者の組織の病域の特定、或いは健常人の遺伝子を解析することでその病気にかかりやすいヒトを選別することができる。そうすることで、患者に対する治療の方法、例えば、その保因者とそうでない人に対しては、例えばポリプ切除の緊急性の判断に応用できる。健常人に関してはその病気に関する診断検査のそれぞれの人の遺伝子情報に基づき使い分けることが可能である。

【0060】

また、相同性比較の対象とするデータ群として、遺伝子以外のデータを用いることにより、人材情報データの解析、選別や、商品の類似性判定、気象情報解析

などにも応用できる。すなわち、複数のデータからなるデータ群同士の類似性の度合いを解析するものであれば何でも適用可能である。

【 0 0 6 1 】

【発明の効果】

以上詳述したように本発明によれば、多数の遺伝子同士を定量的にかつ正確に比較することが可能となる。

【図面の簡単な説明】

【図 1】

本発明の第 1 実施形態に係る遺伝子解析システムの全体構成を示す図。

【図 2】

同実施形態に係る遺伝子解析に必要な遺伝子データを取得するプロセスの一例を示す図。

【図 3】

同実施形態に係るアミノ酸配列記憶手段に格納された遺伝子データ群の概念図。

【図 4】

同実施形態に係る B L A S T 法の概念図。

【図 5】

同実施形態に係る B L A S T 法による相同性の有無の判定の概念を示す図。

【図 6】

同実施形態に係る相同値ヒット数としきい値 E との関係を示す図。

【図 7】

同実施形態に係る相同性決定値 $Z_i(1)$ を示す図。

【図 8】

同実施形態に係る T 検定における判定結果導出に用いられる判定テーブルの一例を示す図。

【図 9】

同実施形態に係る T 検定に於ける判定結果により考察した酵母遺伝子の由来を模式的に示す図。

【符号の説明】

1…プロセッサ

2…出力手段

3…アミノ酸配列記憶手段

4…相同値記憶手段

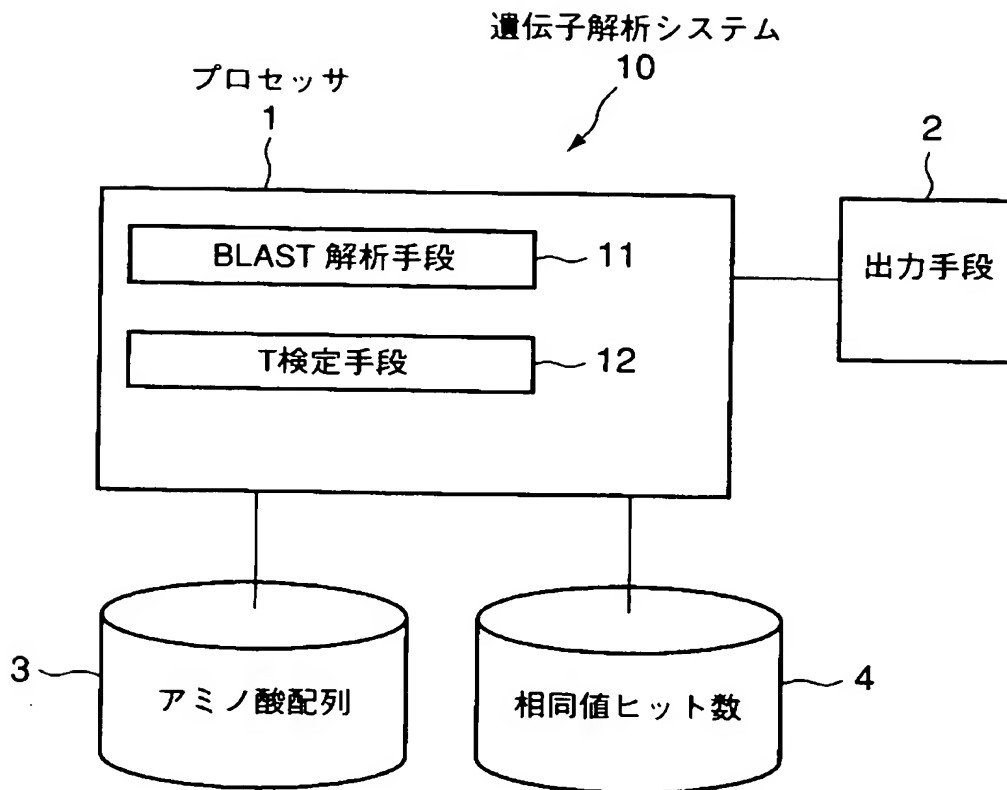
1 1…B L A S T解析手段

1 2…T検定手段

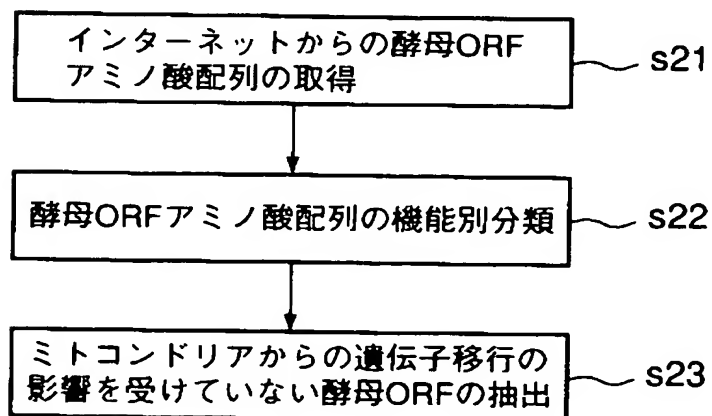
【書類名】

図面

【図 1】



【図 2】



【図 3】

酵母遺伝子データ					
ORF1	ORF2	ORF3	ORF4	ORF5	ORFp
アミノ酸a,b,c	アミノ酸a,b,d	アミノ酸c,d,e	アミノ酸f,g,h	アミノ酸z,r,t,g,h	アミノ酸e,f,g,b,x

古細菌A遺伝子データ					
ORF1	ORF2	ORF3	ORF4	ORF5	ORFn1
アミノ酸d,y,e	アミノ酸w,x,b	アミノ酸x,d,b	アミノ酸w,x,d	アミノ酸s,d,h	アミノ酸x,f,y

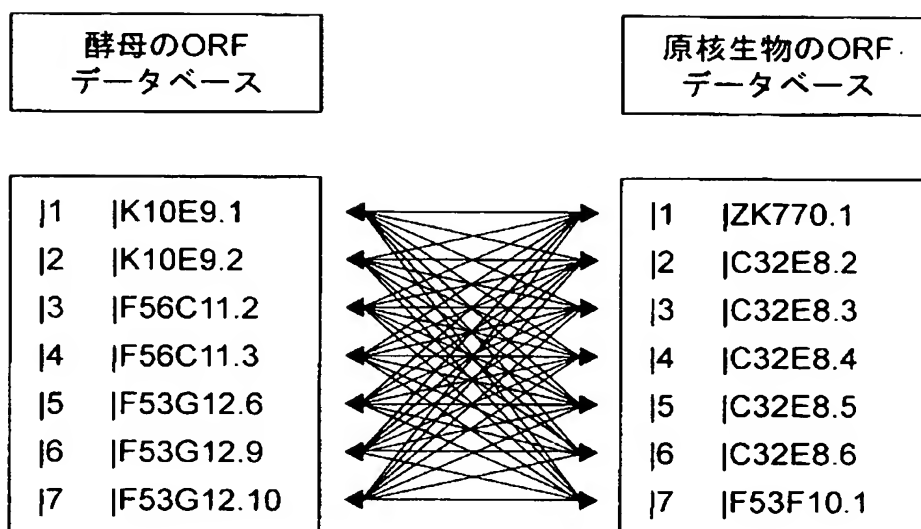
古細菌B遺伝子データ					
ORF1	ORF2	ORF3	ORF4	ORF5	ORFn2
アミノ酸d,e,s	アミノ酸z,d,s	アミノ酸x,b,j	アミノ酸b,y,i	アミノ酸p,b,c	アミノ酸t,y,x

:

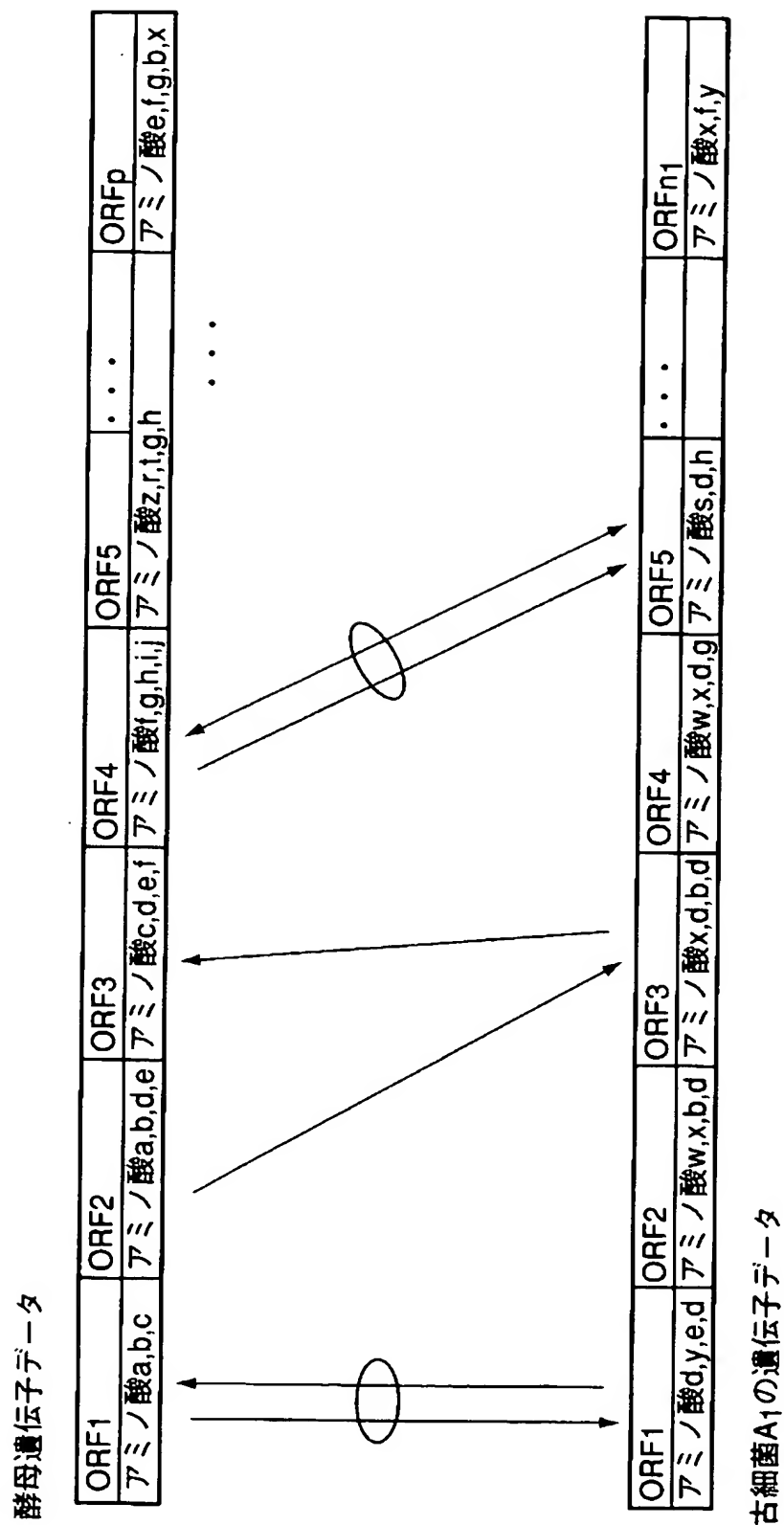
真性細菌C遺伝子データ					
ORF1	ORF2	ORF3	ORF4	ORF5	ORFn3
アミノ酸t,b,g	アミノ酸x,f	アミノ酸q,c	アミノ酸n,m	アミノ酸x,l	アミノ酸w,r

真性細菌D遺伝子データ					
ORF1	ORF2	ORF3	ORF4	ORF5	ORFn4
アミノ酸y,c	アミノ酸w,e,r	アミノ酸q,x,t	アミノ酸p,b	アミノ酸q,n,c	アミノ酸o,v,u

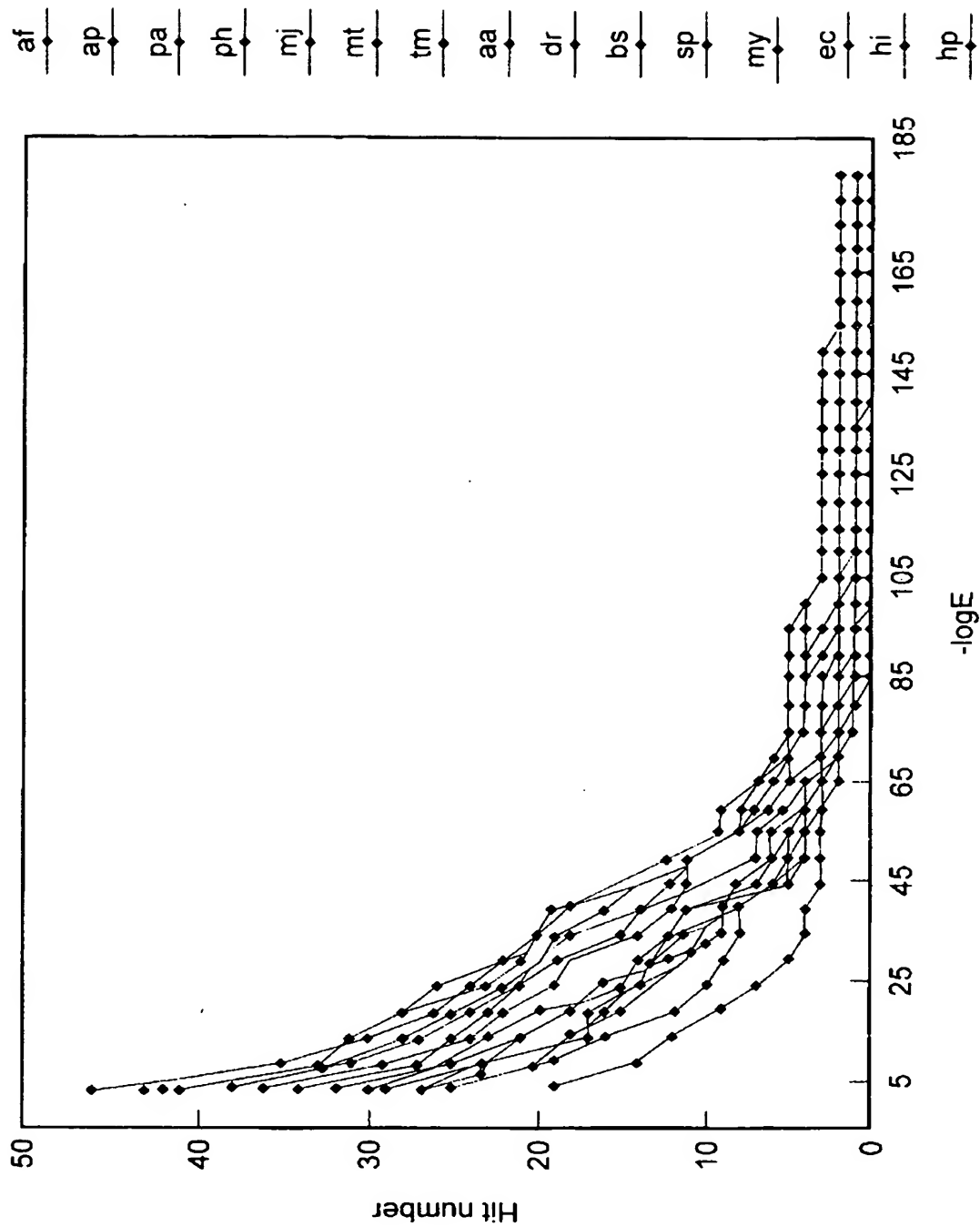
【図 4】



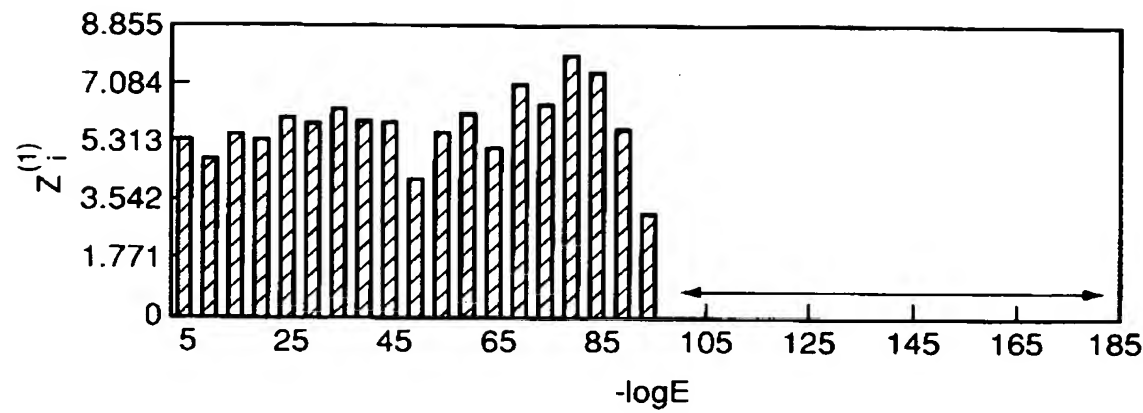
【図 5】



【図 6】



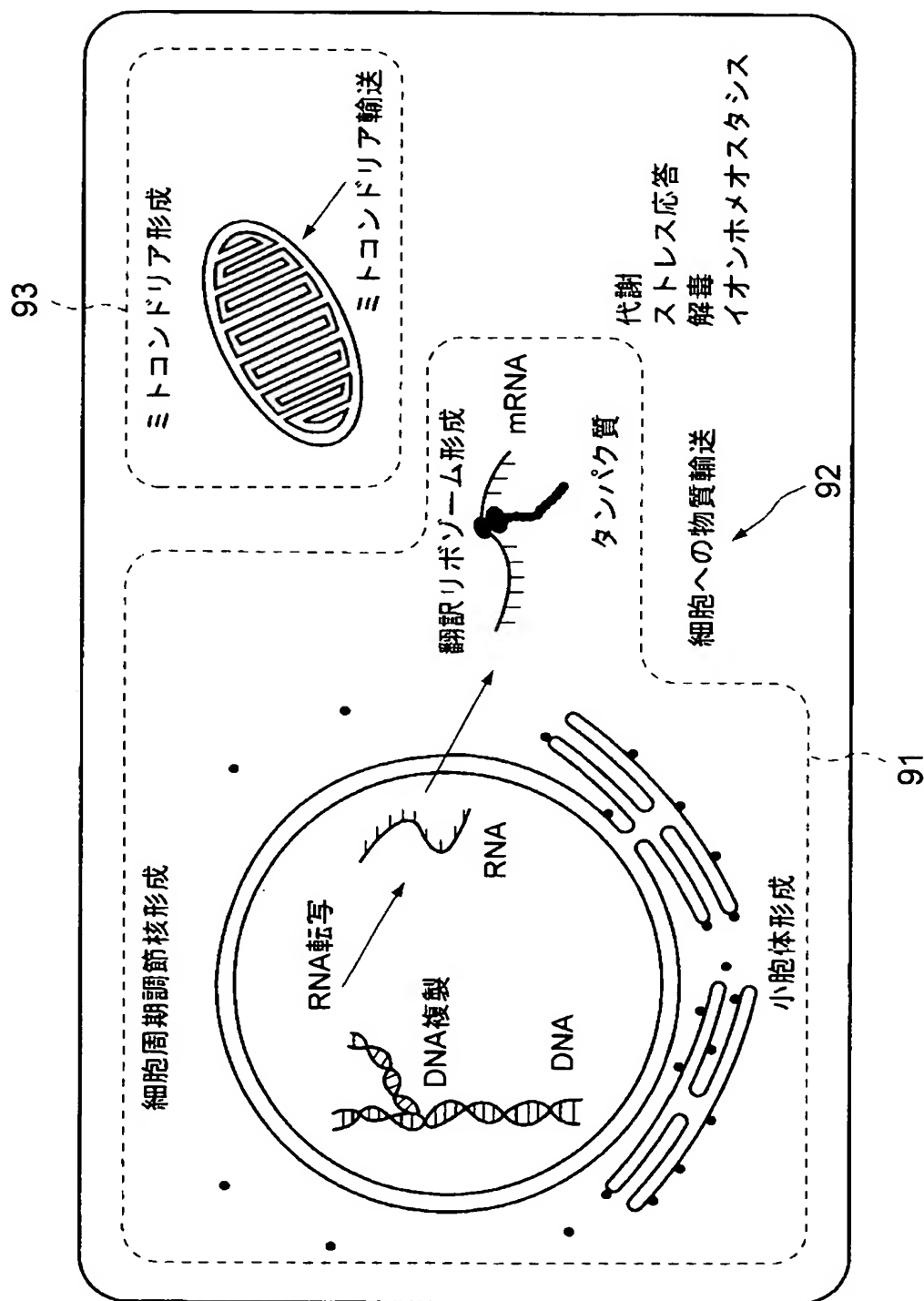
【図 7】



【図 8】

$Z^{(2)}/t_{n-1} \backslash \overline{Z^{(1)}}$	+	-
1以上	古細菌	真正細菌
1未満	判定不可	判定不可

【図 9】



【書類名】 要約書

【要約】

【課題】 多数の遺伝子同士を定量的にかつ正確に比較する。

【解決手段】 B L A S T解析手段 1 1により、解析対象データ群と第 1 のデータ群のそれぞれのデータ群に含まれるデータの相同性を計算し、その値を用いて相同性有りとなるデータの数（オルソログス遺伝子数）を第 1 の相同値 x として算出し、しきい値 E を n_A 個設定して各しきい値 E_i 毎に第 1 の相同値 x_i を算出し、解析対象データ群と第 2 のデータ群のそれぞれのデータ群に含まれるデータの相同性を計算し、相同性有りとなるデータの数（オルソログス遺伝子数）を第 2 の相同値 y として算出し、しきい値 E を n_B 個設定して各しきい値 E_i 毎に第 2 の相同値 y_i を算出し、T 検定手段 1 2 は、第 1 の相同値 x_i 及び第 2 の相同値 y_i 並びにしきい値の数 n との関係に基づいて、解析対象データ群が第 1 のデータ群又は第 2 のデータ群のいずれに類似するかを判定する。

【選択図】 図 1

特願 2001-183856

出 願 人 履 歴 情 報

識別番号

[396020800]

1. 変更年月日

1998年 2月24日

[変更理由]

名称変更

住 所

埼玉県川口市本町4丁目1番8号

氏 名

科学技術振興事業団